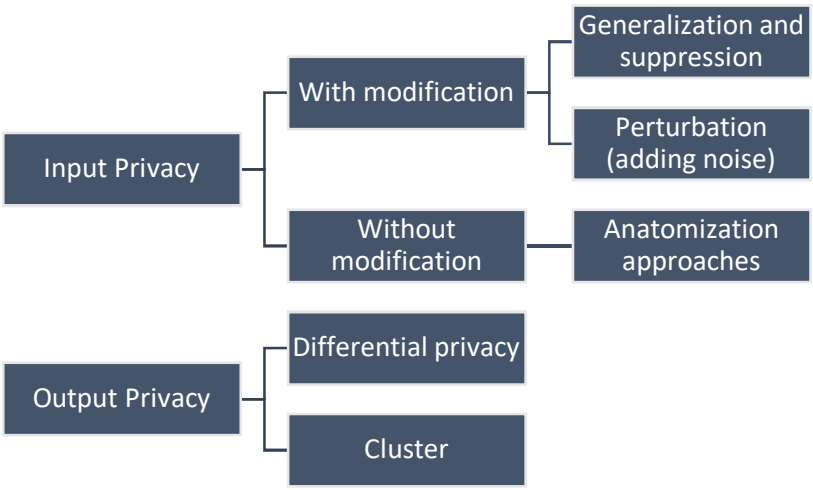
Weekly report

1 Done

1.1 Reading:

1.1.1 **Opportunities and Challenges for Privacy-Preserving Visualization of Electronic Health Record Data** (Dasgupta):



- Visual representations and the associated interaction mechanisms both have their own challenges in protecting unintended disclosure.
- The evaluation of privacy risks should focus on "the amount of labor required to" break privacy.

1.1.2 **Measuring Privacy and Utility in Privacy-Preserving Visualization** (Dasgupta)

- Several metrics proposed to quantify privacy and utility individually:
  - Entropy as a privacy metric.
  - Data quality and clustering quality as utility metrics.
- Uncertainty is one of the keys to balance effective privacy-preservation and analysis.
- Visual uncertainty can be decomposed into a set of encoding and decoding uncertainties

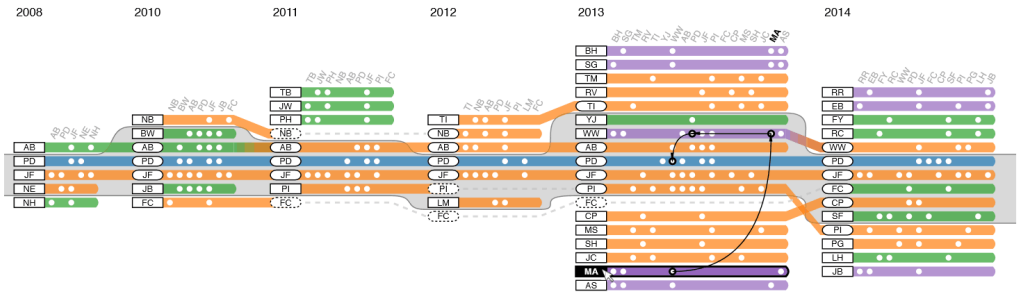
	Cause/Effect of Uncertainty	Measurable Quantities	Measured Criteria
Encoding	Precision Granularity	Binning, Cluster range	Privacy Privacy
Decoding	Spatial Accuracy Identity Traceability Pattern Complexity	Cluster Range Cluster Overlaps Cluster Splits Semantic Structures	Privacy Privacy, Utility Privacy, Utility Utility

- Metrics for uncertainty: cluster summary error, cluster range, overlap

clutter, overlap entropy, mutual information and average split count.

1.1.3 Egocentric Analysis of Dynamic Networks with EgoLines (Jian Zhao)

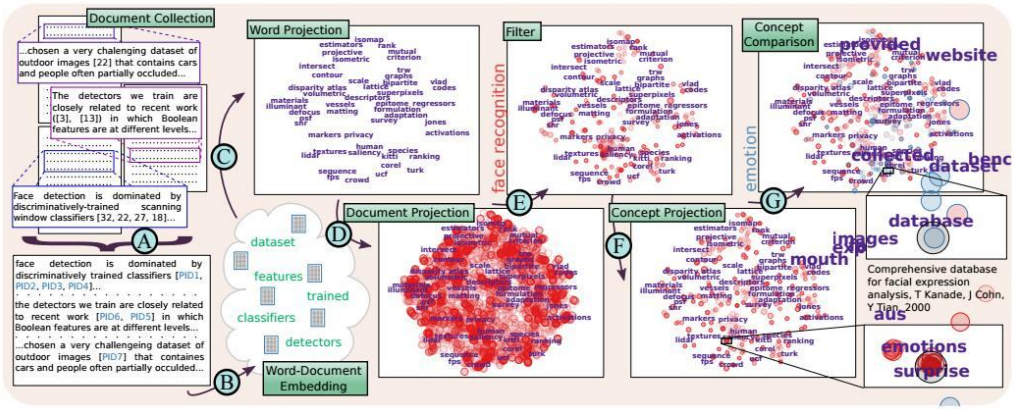
EgoLines is an interactive visualization using a "subway map" metaphor to support the egocentric analysis of dynamic networks.
The correlations are visualized as white circles in the lines, which compose one matrixes for a period of time.



This kind of visualization approach does not involve privacy issues. If we design some icons to represent geographic information or event information, I think it can be applied in RelationLines as the main view. However, the interactions need to be well-designed so that previous analysis can be kept.

1.1.4 cite2vec: Citation-Driven Document Exploration via Word Embeddings (Matthew Berger)

This paper introduces a novel method for modeling and exploring documents via their citation contexts.

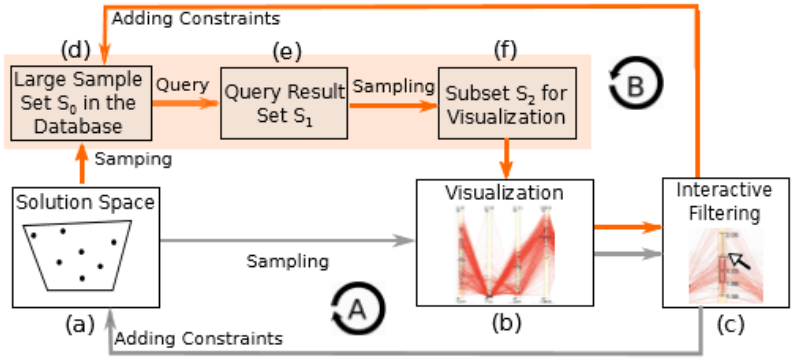


From a given document collection they resolve all document citations to unique identifiers, and jointly learn a semantic embedding of words and documents. They then perform a 2D projection of the words and documents from the embedding, and allow the steering of document projections via user-defined concepts. Here, the user specifies “face recognition”, resulting in each

word being composed with this provided phrase. Upon specifying “emotion” we observe a document change its projection due to a better word-phrase composition.
- Skip-gram Model: aiming to find an embedding such that words which co-occur are close in the embedding.
- Words for projections: covering general themes found within documents; semantically distinct; reduced by sampling.

1.1.5 A Visual Analytics Approach for Categorical Joint Distribution Reconstruction from Marginal Projections (Cong Xie)

This paper aims at reconstructing the joint distribution from a multi-set of marginal distributions. Their system combines domain knowledge from experts and data feature. Boxplot and heatmap are employed to visualized probability distribution. They choose line chart to compare the solution space before and after setting constraints. They re-sample the solution space when changes happened.



1.2 Learning (Differential Privacy)

https://research.neustar.biz/2014/09/08/differential-privacy-the-basics/
- Definition: A randomized function K gives ε-differential privacy if for all data sets D and D′ differing on at most one row, and all S ⊆ Range(K),

$$Pr[K(D) \in S] \leq exp(\varepsilon) \times Pr[K(D') \in S]$$,

which means that the risk to one’s privacy should not substantially (as bounded by ε) increase as a result of participating in a statistical database.
- Mechanism design:

* Laplace mechanism: involves adding random noise that conforms to the Laplace statistical distribution. The 0-centered Laplace distribution has only one parameter (its scale), and this is directly proportional to its standard deviation, or noisiness. We define the parameter depending on the privacy parameter, ε and the query, $\Delta f = \max_{D,D'} ||f(D) - f(D')||_1$:

$$\Delta f / \varepsilon$$

* Exponential mechanism: is more appropriate for categorical data.

1.3 Project (map information)

1.3.1 Set up the framework of our program.

1.3.2 Tianyi focused on the part of heatmap, Wenlong learnt js and our previous code, and Zhengshang kept solving the layered road network.

1.4 Project (differential privacy)

1.4.1 Learnt materials they gave.

1.4.2 Had a talk with Tianyi and sent our idea to the group.

1.5 Others

1.5.1 Presented paper at group meeting.

1.5.2 Wrote related blog.